Quality and User Experience (2018) 3:6
https://doi.org/10.1007/s41233-018-0019-8

RESEARCH ARTICLE

Interpreting MOS scores, when can users see a difference? Understanding user experience differences for photo quality

Stamos Katsigiannis[1] · James Scovell[2] · Naeem Ramzan[1] · Lucjan Janowski[3] · Philip Corriveau[2] · Michele A. Saad[4] · Glenn Van Wallendael[5]

Received: 12 September 2017 / Published online: 2 May 2018

Abstract
The use of no-reference image quality evaluation tools that produce MOS scores, like the VIQET tool which was released by the Video Quality Expert Group, raises the question of whether the produced MOS differences between images correspond to noticeable differences in quality by the consumers. In this work, we attempted to approximate the minimum MOS difference that is required in order for people to be able to distinguish between a higher and a lower quality image under realistic conditions that are commonly encountered in the current consumer space. 91 people participated in a subjective just-noticeable-differences study across three countries that used non-simulated image stimuli, produced and evaluated through crowd sourcing for the validation of the VIQET no-reference image quality tool. The image dataset consisted of 15 different scenes belonging to three different scene types, with a total of 210 different image pairs being used. After evaluating the quality of the collected data, a logistic regression analysis approach was employed in order to estimate the minimum MOS difference required between two images in order for a given percentage of people to be able to detect the higher quality image.

Keywords QoE · MOS difference · JND · Just noticeable difference · Image quality · VIQET

Introduction

For several decades, the media industry and the image quality research community have worked on developing and deploying no-reference image quality evaluation tools [10]. This challenging problem requires developing prediction models that algorithmically map photos to scores representative of human judgments of perceived image quality. A typical measure of perceived image quality is known as the Mean Opinion Score (MOS). MOS is obtained by asking observers to rate images for their quality on a particular scale (such as a scale from 1 to 5 where 1 is bad and 5 is excellent). Image quality evaluation tools hence serve to algorithmically predict image MOS that would otherwise have been obtained by asking a number of observers their opinion about the quality of

✉ Stamos Katsigiannis
Stamos.Katsigiannis@uws.ac.uk

James Scovell
James.J.Scovell@intel.com

Naeem Ramzan
Naeem.Ramzan@uws.ac.uk

Lucjan Janowski
janowski@kt.agh.edu.pl

Philip Corriveau
Philip.J.Corriveau@intel.com

Michele A. Saad
Michele.A.Saad@intel.com

Glenn Van Wallendael
Glenn.VanWallendael@ugent.be

[1] School of Engineering and Computing, University of the West of Scotland, High St, Paisley PA1 2BE, UK

[2] Intel Corporation, 2111 NE 25th Ave, Hillsboro, OR 97124, USA

[3] Department of Telecommunication, AGH University of Science and Technology, Kraków, Poland

[4] Intel Corporation, 1300 S. Mopac Exp., Austin, TX 78746, USA

[5] Department of Electronics and Information Systems, Ghent University, Ghent, Belgium

an image. An example of a no-reference tool for image quality prediction is the Video Quality Expert Group's tool called VQEG Image Quality Evaluation Tool (VIQET) [18] which was released in March 2016 and open-sourced to the community.

A question that arises from using a tool like VIQET that produces MOS ratings is whether the MOS differences produced are noticeable to a consumer. For instance, on a five point MOS scale, an image that scores a MOS value of 5 is expected to be noticeably better in quality than one that scores a MOS value of 1. On the other hand, it is unclear whether an image that scores a MOS score of 3.8 is noticeably better in quality than one that scores a MOS score of 3.6. Consequently, it is not yet clear which is the smallest MOS difference that is perceptible to users.

In this work, we present a subjective study that seeks to determine the smallest noticeable MOS difference on a set of images that are representative of consumer usage scenarios, i.e. using image examples that are commonly encountered in the current consumer space. Image quality expectation, and hence MOS, change over time as consumer expectations change. As a result, the minimum noticeable difference determined in this work may change over time. However, the main contribution of this work is the methodology and subjective study approach we have taken to answer the question of perceptible quality difference. We refer to our subjective study as a noticeable-differences subjective study. The proposed approach is reminiscent of traditional just-noticeable-differences (JND) approaches [23, 30], but deviates from them in that it only uses non-simulated image stimuli, i.e. images that have not been artificially degraded or altered in any way.

Participants were presented with pairs of images and were asked to select the higher quality image. Regression analysis was then employed in order to map the success in selecting the higher quality image to the difference in MOS between two images, thus the minimum MOS difference in order for a given percentage of people to be able to detect the higher quality image was established. Furthermore, in order to take into consideration possible differences in quality perception due to cultural background and location, a cross-laboratory verification was conducted by using data gathered from three labs located in three different countries (Belgium, UK, USA).

The rest of this paper is organized in five sections. "Background" section provides background information on the issue of just-noticeable-differences for image quality assessment and how it is approached in this work. The methodology followed is described in "Methodology" section, while results are analyzed and evaluated in "Results" section, and discussed in "Discussion" section. Finally, conclusions are drawn in "Conclusion" section.

Background

In psychophysics, the JND is the smallest delta that a stimulus needs to change before it is perceived by at least 50% of the subjects that are presented with the stimulus [12]. JND studies have found applications in various areas of psychophysics including sound perception such as music and speech [27], visual perception [2, 11], and haptics [8, 12, 27]. In this work, we narrow the focus of our attention to visual perception, specifically the perception of image quality.

According to Wu et al. [23], JND models for visual perception can be divided into two categories based on the domain used for computing the JND threshold. These two JND categories are the subband-domain JND models and the pixel-wise JND models. For the subband-domain models, the image needs to be first transformed into a subband domain, such as the DCT-domain [9, 21]. On the other hand, pixel-domain methods [13, 19, 24, 26, 29] are directly calculated on the spatial domain and thus are more convenient and less computationally complex. As a result, they have been used for a wide variety of applications such as visual quality assessment and enhancement [26].

Various JND models have been proposed in the literature. Yang et al. [25, 26] proposed the use of the overlapping effect of luminance adaptation and spatial contrast masking in order to create a JND model, while Wu et al. [22] proposed a JND estimation model based on the free-energy principle unified brain theory. Ahumada and Peterson [1] developed a well-cited DCT scheme for JND based on the spatial contrast sensitivity function (CSF). Many later works were based on this scheme in order to provide simpler or more advanced JND models [4, 17, 20, 28]. Peterson et al. [14] extended this scheme for color images, while Watson [20] improved it into the DCTune model by taking contrast masking into consideration. Hontsch and Karam [4] further modified the DCTune model by considering a foveal image region instead of only single pixels and used a locally adaptive perceptual quantization scheme based on a tractable perceptual distortion metric.

Methodology

Traditional image quality JND studies follow a design methodology that requires the generation of the image stimuli artificially. The traditional design, takes an original reference image and introduces a particular adjustment to it (such as the introduction of noise or blur or compression) at incrementally increasing levels of severity [3, 12]. The approach we propose here deviates from the

traditional JND study design in that it relies solely on non-simulated image stimuli. The study utilizes only images produced by a multitude of consumer devices, namely phones, tablets, compact cameras, and digital single-lens reflex (DSLR) cameras. The reason behind this non-conventional approach is that no-reference image quality evaluation research efforts have shifted focus to non-simulated image stimuli in order to better model the overall impact of a capture system on perception and human quality judgment. Given the difficulty in simulating the joint impact of optics (lens and sensor) and post processing (typically a non-disclosed black box that varies widely between consumer devices), our dataset of image stimuli for the noticeable-difference study we describe in this work is a set of photographs of multiple scenes captured by a multitude of devices and hence exhibiting disparate image quality.

Environment and participants

Labs in Belgium (*UGhent*), United Kingdom (*UWS*), and the United States (*Intel*) were used for this study (Table 1). Participants were seated in an ergonomic chair in front of a desk with two monitors in front of them. A keyboard and mouse was used to make selections. The distance between the displays and participants was the standard three times the height of the display [5]. While the displays varied between labs, the two displays in each lab were identical. *Intel* and *UGhent* utilized two Samsung 28″ UHD monitors (3840 × 2160), model no. U28D590D, while two Sony Bravia 55″ 4K TVs (3840 × 2160), model no. XD93, were used by *UWS*. The screens were calibrated according to Rec. ITU-R 709 [7] using an Atomos Spyder color calibration unit to sRGB gamut, D65 white point (6500K), 120 cd/m^2 brightness, and minimum black level. Furthermore, the color of the walls or curtains present in the test area was mid gray.

A total of 91 participants completed the study: 36 participants (19 male, 17 female) at the United States lab (*Intel*), 31 (27 male, 4 female) at the Belgium lab (*UGhent*), and 24 (19 male, 5 female) at the United Kingdom lab (*UWS*). All participants had normal or corrected eyesight and did not report having any problem with their vision. In total, 65 males and 26 females participated in the study, with their age varying between 20 and 59 years old. Participants from *UGhent* and *UWS* were mostly PhD students and researchers, while participants from *Intel* were employed in various sectors and were recruited through a third party. Furthermore, the average duration of each session across all participants was 26.88 (± 10.70) min. Details about the labs and the participants of this study are summarized in Table 1.

Images

The images used in this study were originally taken in an effort to develop VIQET, a no reference image quality tool [18]. These images were captured by taking photos of the same scenes with a wide range of cameras of various quality (phones, tablets, compact cameras, and DSLR cameras) and are part of the Consumer Content Resolution and Image Quality (CCRIQ) dataset [15], which is available from the Consumer Digital Video Library (CDVL) (http://www.cdvl.org/). The CCRIQ dataset contains 18 different scenes, belonging to 5 topic categories, namely flat surfaces, landmarks at night, landscapes with good lighting, portraits, and still lifes. Out of these 18 scenes, 15 individual scenes were selected for this study as follows: scenes were divided into three scene types, namely indoor, landscape, and night shots, and 5 scenes belonging to each scene type were selected for the experiments in this work.

A sample image from each selected scene is shown in Fig. 1a–o. For each of the individual scenes, fourteen image pairs were selected for Phase 1, totaling 210 photo pairs. Ten of these pairs were repeated as a measure of reliability. As these images had been part of previous research [18], each image had a MOS assigned via crowd sourcing [16] which represented end-user ratings of image quality on a five point scale. MOS is an ITU standard for measurement of subjective assessment [6]. Image pairs were chosen to ensure that within each scene, the MOS deltas between image pairs ranged from small to large ($\delta_{MOS} \in [0.25, 0.95]$). An effort was also made to ensure that image pairs were distributed from low quality to high. For example if there was a 0.4 MOS delta between an image pair of lower quality then there was another pair of images from the same scene that had a delta of roughly 0.4 MOS that was of higher quality. The

Table 1 Labs participating in this study and participant demographics

Lab	Country	Institution	Monitor (model)	Participants		
				All (M/F)	Avg. age (SD)	Occupation
Intel	USA	©Intel Corporation	Samsung 28″ UHD (U28D590D)	36 (19/17)	39.3 (11.9)	Various sectors
UGhent	Belgium	Ghent University	Samsung 28″ UHD (U28D590D)	31 (27/4)	28.0 (7.0)	Ph.D. students/researchers
UWS	UK	University of the West of Scotland	Sony Bravia 55″ 4K (XD93)	24 (19/5)	30.0 (5.7)	Ph.D. students/Researchers

M male, *F* female, *SD* standard deviation

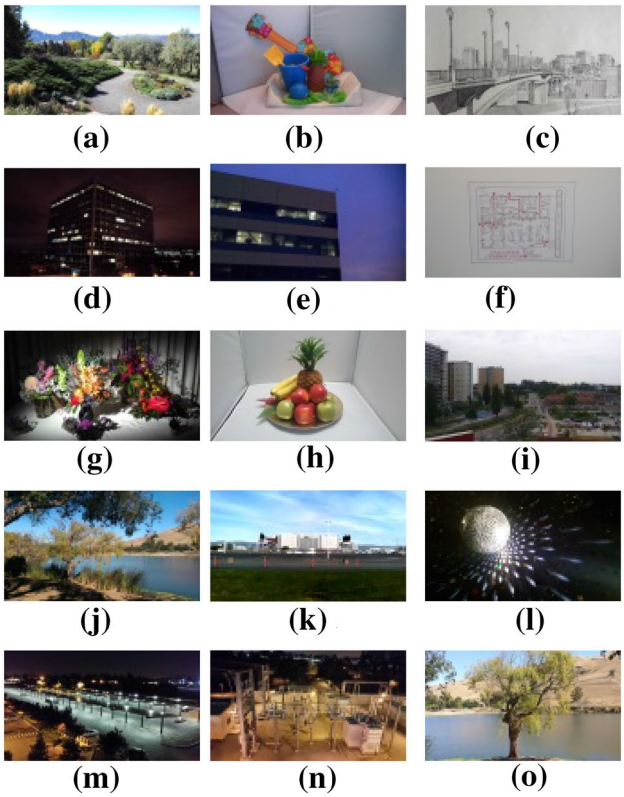

Fig. 1 Sample images from each scene. **a** AutumnMtn, **b** beach toys, **c** bridge, **d** building, **e** build. corner, **f** evac. plan, **g** flowers, **h** fruit, **i** Ghent, **j** green tree, **k** Levi, **l** mirror ball, **m** parking, **n** pipes, **o** tree lake

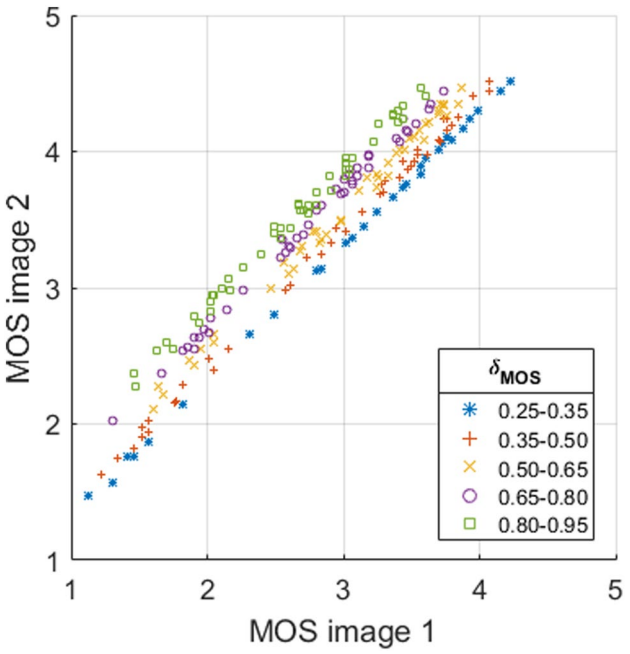

Fig. 2 Distribution of delta MOS values in relation to the absolute MOS values of the images in each pair

distribution of MOS deltas in the dataset in relation to the absolute MOS values of the images in each pair is shown in Fig. 2.

For Phase 2, 51 images from the ones used in Phase 1 were selected for each lab as follows: For each lab, 3 individual images were selected from each of the 15 scenes used in Phase 1, leading to a total of 45 images. Out of these 45 images, 9 were selected as control images and were the same for all labs and 36 were unique to each lab. Furthermore, 6 out of the 9 control images were repeated for each lab, leading to 15 control images that were similar for all the labs. Of the 15 control images, the unique 9 were designed to act as a measure of reliability across labs and the 6 repeats were designed to act as a measure of participant reliability. As a result, 15 control and 36 unique images were selected for each lab, leading to a total of 51 images for phase 2. Similar to Phase 1, an effort was made to span the range of MOS scores for the photos that were rated across all three labs.

Procedures

A software application was designed to automate the sessions and guide participants through the entire study. There were two phases to this study which all participants completed. Phase 1 was designed to measure how large of a difference in image quality there needed to be for participants to determine the image of higher quality. To this end, participants were asked to select the image they considered as having the higher quality between two given images. Phase 2 was designed to ensure that participant ratings of images was not significantly different from previous research. In that phase, participants were asked to rate each given image according to its quality. To avoid any misunderstanding or different interpretation of what *high quality* means, participants were instructed that overall image quality refers to how good the picture looks, taking into consideration the entire aspects of the picture (color, crispness, detail). Furthermore, they were asked to keep in mind that they were not rating whether they like the content, whether they like the people in the picture, or whether they like the composition of the picture. Apart from the actual experiment session, each phase included a short practice session to ensure participants understood the task. The practice session for each phase consisted of the same task as the actual experiment, but using only five pairs of sample images and five sample images for Phase 1 and 2 respectively.

Phase 1 was designed to present a full screen view of a photo on each monitor. The images were of the same scene but were taken with different cameras to ensure photo quality variation, so images were not the exact same frame. Participants selected the image on the left or right display by hitting the corresponding left or right arrow followed by "Enter" on the keyboard. When a participant made a selection, the

software automatically changed to the next image pair and continued until all 220 image pairs had been selected. Photo pairs were presented in a random order although the software was designed to never present the same scene more than three times in a row. The software also randomized the images between the two displays. The same 220 image pairs were presented to all 91 users.

Phase 2 began with a practice session which started automatically when Phase 1 ended. Phase 2 presented a single full screen image on the left display and a five point ratings scale (5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad) on the right display. Participants were asked to rate the "image quality" on the left display. When a participant made a selection, the software automatically moved on to the next image until all 51 images were rated. Images were presented in a random order.

Phase 2 was critical because if the images were rated significantly different from the original MOS ratings collected through a crowd sourcing study then the deltas that were used to select image pairs, as described above, would not be accurate.

Results

In order to extract meaningful and safe conclusions, the data obtained in this study were first evaluated in terms of the recorded MOS, agreements between different labs, screen selection, and scene type.

Evaluation of captured data in terms of MOS

The goal of this study is to estimate a function $p = f(\delta_{MOS})$, where p is the probability of selecting an image with higher MOS for the task: "select a better quality image between two given images", and δ_{MOS} is the difference of MOS of the two compared images. The proposed model is built based on the MOS collected through crowd sourcing [16]. Therefore, we have to ensure that MOS collected in the laboratory study correlate well with the MOS collected by the crowd source experiment. As a result, in order to establish the quality of the captured data, the MOS values recorded for each image at the second phase of the experiment by each lab were compared to the MOS values received for the same images through crowd sourcing.

To evaluate their similarity, the Pearson's correlation coefficient (PCC) was computed, indicating a strong ($PCC_{UGhent} = 0.8938$) or very strong correlation ($PCC_{Intel} = 0.9632$, $PCC_{UWS} = 0.9276$) between the crowd sourced MOS and this studies MOS for each of the labs, as well as a very strong correlation for the ratings across all labs ($PCC_{all} = 0.9111$), as shown in Table 2. Four one-way analyses of variance (ANOVAs), each between this studies MOS per image and the crowd sourced MOS per image, for each of the labs, as well as for all the labs together, showed that there is no statistically significant difference ($p > 0.09$ in all cases) between the ratings (detailed results in Table 2). Furthermore, the PCC between each subject's MOS ratings and the crowd sourced MOS is shown in Fig. 3, indicating a subject-wise strong correlation.

Linear fitting of the crowd sourced MOS into the MOS received through this study also showed that there is a good relationship between the ratings, as shown in Table 2 and in Fig. 4. The results from the linear fitting are provided in Table 2 in the form of the resulting linear equations and the R^2, while plots for both the data and the resulting equations are shown in Fig. 4 for each lab separately and in Fig. 5 for all the labs. It must be noted that while the

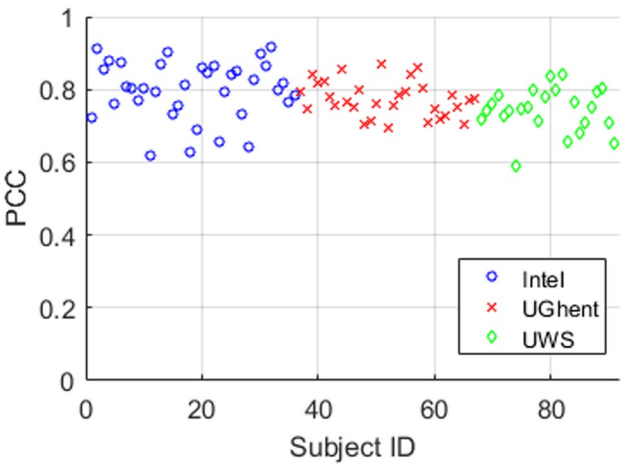

Fig. 3 Pearson's correlation coefficient between the MOS ratings of each subject and the crowd sourced ratings

Table 2 Relationship between the MOS received for each image through this study (y) and through crowd sourcing (x)

Lab	PCC	ANOVA p	Linear fit	CI_β	CI_α	R^2
Intel	0.9632	0.9037	$y = 0.9975x - 0.0190$	± 0.0722	± 0.2246	0.9277
UGhent	0.8938	0.0917	$y = 1.1008x - 0.6665$	± 0.1416	± 0.4357	0.7989
UWS	0.9276	0.7165	$y = 1.0316x - 0.0243$	± 0.1065	± 0.3314	0.8605
All	0.9111	0.3680	$y = 1.0451x - 0.2429$	± 0.0682	± 0.2113	0.4379

CI, 95% confidence interval; β, slope parameter; α, intercept parameter

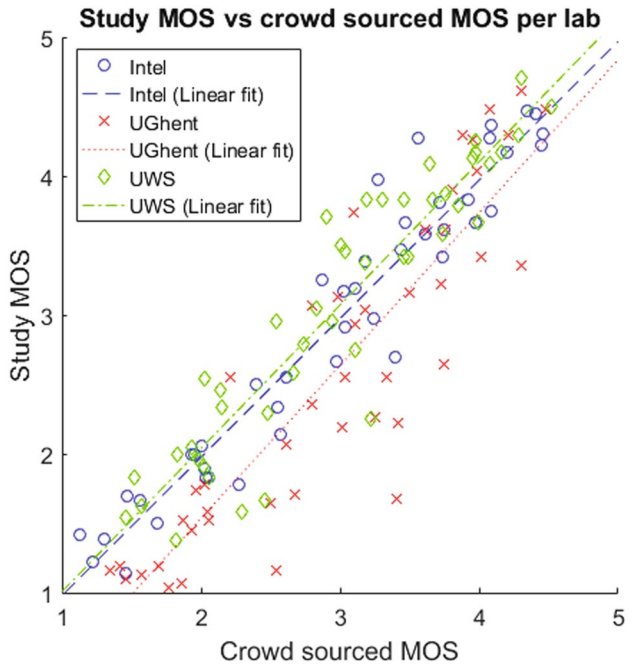

Fig. 4 The MOS received through this study compared to crowd sourced MOS ratings, along with their linear fitting, for each lab participating in the study

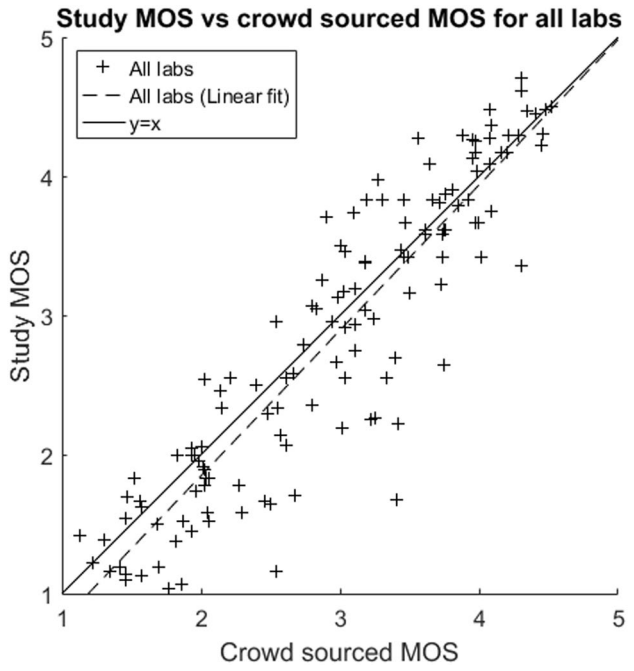

Fig. 5 The MOS received through this study compared to crowd sourced MOS ratings, along with their linear fitting, for all the labs participating in the study

ratings received from *UGhent* are lower on average than the other two groups, there still exists a good relationship between the MOS from *UGhent* and the crowd sourced

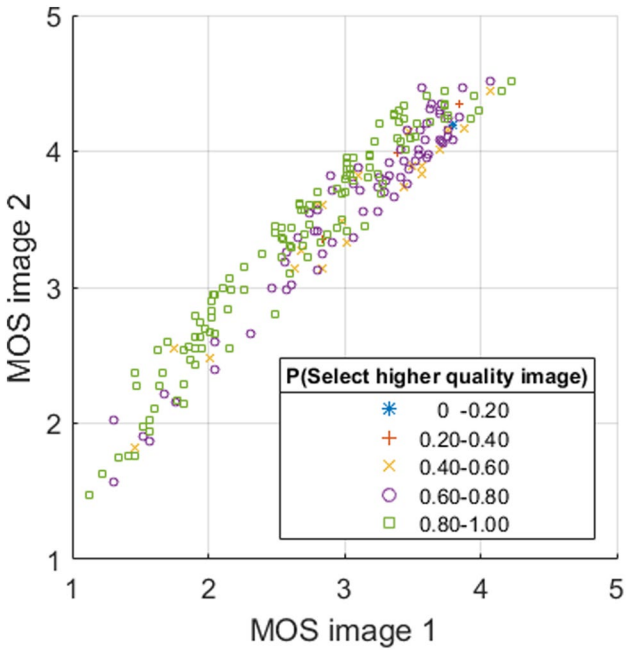

Fig. 6 The percentage of participants that successfully selected the higher quality image among a pair of images, in relation to the MOS rating of each image

data [PCC = 0.8938, no statistically significant difference ($p = 0.0917$), $R^2 = 0.7989$].

Apart from the relation of the acquired MOS ratings from each lab to the MOS ratings from the crowd sourcing experiments, it is important to establish whether the range of the MOS values of the images in a pair affect the ability of the users to detect the higher quality one. To this end, the percentage of the participants that successfully detected the higher quality image from each pair was computed and plotted in Fig. 6 against the MOS ratings of the images in the pair. From this figure, it is evident that high success rates are almost evenly spread across all ranges of MOS values, thus no relation between the range of the MOS ratings and the ability of the users to select the higher quality image can be established. An interesting observation is that most of the cases of low or medium success rates refer to pairs with MOS ratings above 2.5. Nevertheless, the vast majority of pairs with MOS ratings above 2.5 had success rates above 0.60. This shows that while some users had some difficulty when comparing images with high MOS values, the majority of the users were able to detect the higher quality image regardless of the range of the MOS values.

Evaluation of captured data in terms of consistency across labs

The positioning of the higher quality image across the two screens for each similar pair of images differed between

participants, i.e. for the same pair of images, some participants would have the higher quality image reside on the left screen, while others would have the higher quality image reside on the right screen. For each pair of images, approximately 50% of the participants had the higher quality image reside on the left screen and approximately 50% on the right screen. After computing the percentage of participants that detected the higher quality image for each image pair when they selected the left screen and when they selected the right screen, and in order to establish the quality of the captured data, a correlation analysis between the success rates per selected screen for each image pair and for each lab was conducted. The PCC computed between the average success rate for each image pair for when the right or the left screen was selected at each lab is shown in Table 3. It would be expected that the difference in the success rates for the same pairs of images when the participants selected a different screen would be minimal, and thus the results between different screen selections would be strongly correlated. The results from *UGhent* exhibited a strong correlation between left and right screen selection, whereas a moderate correlation was observed for the results from *Intel* and *UWS*.

In order to further examine the lower correlation for *Intel* and *UWS* between the results of the left and right screens for the same pairs of photos, a one-way ANOVA between the average success rate for each image pair for when the right or the left screen was selected at each lab was conducted. Results showed that there was a statistically significant difference ($p = 0.0109$) between the left and right screen results for Intel, there was no statistically significant difference ($p = 0.3349$) between the left and right screen results for *UGhent* and there was a statistically significant difference ($p = 0.0045$) between the left and right screen results for *UWS*. These findings are consistent with the correlation analysis above and further demonstrate the variation between the results for the same pairs of photos for *Intel* and *UWS*, when the higher quality photo is on a different screen.

Apart from the expected similarity between the success rates for the same image pairs when the higher quality image is on a different screen, it would be expected that the overall success rates (without examining the screen that the higher quality image resided) between labs would not differ significantly, since given a sufficient number of participants in the study, the perception of image quality should statistically be similar. In order to investigate this argument, an ANOVA between the overall success rates per image pair for the three labs was conducted, resulting to a statistically significant difference ($p = 0.0111$). The ANOVA results are also shown in Fig. 7 in the form of box plots.

The same procedure was then repeated twice; once for the average success rate per image pair for when the left screen of each lab was selected and once for the average success rate per image pair for when the right screen was selected. Results are shown in the form of box plots in Figs. 8 and 9 for when the right and left screen was selected respectively, and showed that there was no statistically significant difference ($p = 0.8715$) between the results of the three labs

Table 3 Correlation between the results from when the left or the right screen was selected at each lab for the same pairs of images

Lab	PCC	Correlation	ANOVA p
Intel	0.686	Moderate	0.0109*
UGhent	0.821	Strong	0.3349
UWS	0.624	Moderate	0.0045*

*Statistically significant difference

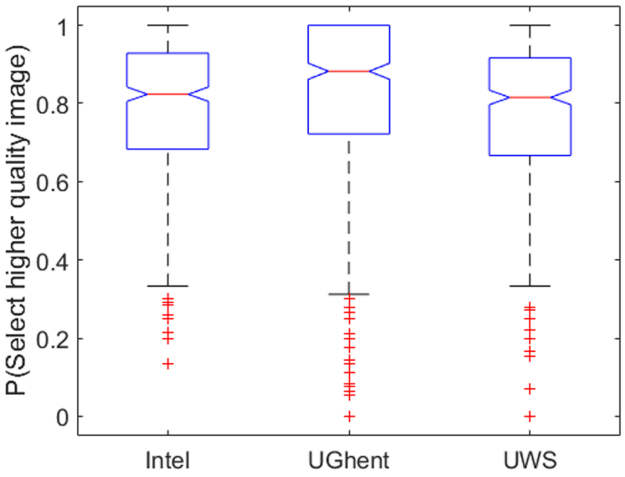

Fig. 7 One-way ANOVA results between all the results (both left and right screens) of each lab

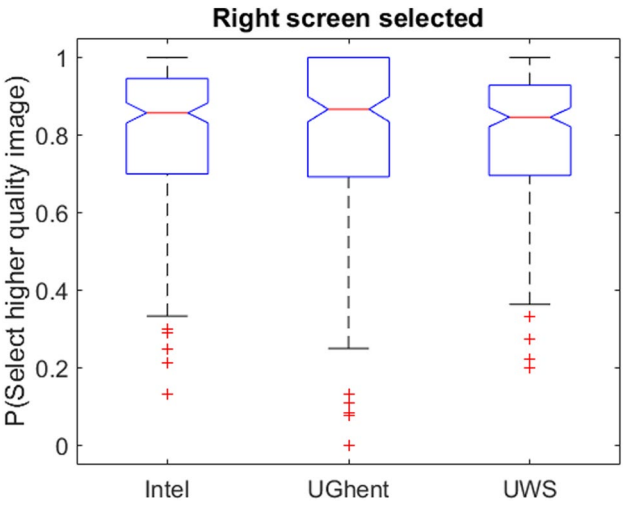

Fig. 8 One-way ANOVA results between the results from the right screen of each lab

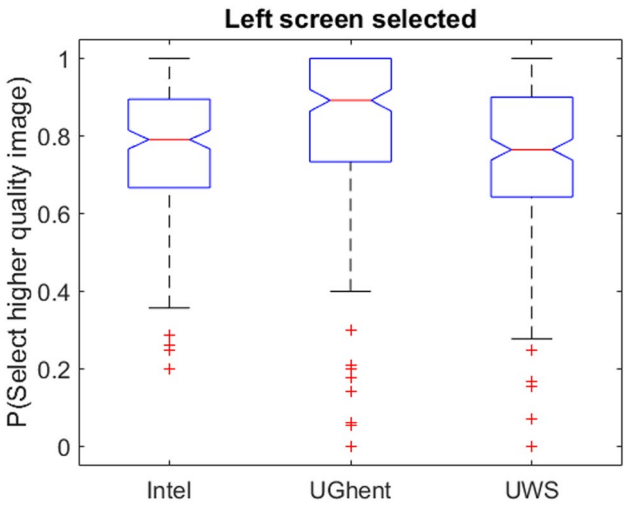

Fig. 9 One-way ANOVA results between the results from the left screen of each lab

Table 4 Results of the one-way ANOVA (p values) between the results of each lab (for both left and right screens)

Lab	Intel	UGhent	UWS
Intel	–	0.0595	0.2618
UGhent	0.0595	–	0.0046*
UWS	0.2618	0.0046*	–

*Statistically significant difference between the two results

when participants selected the right screen. Nevertheless, the ANOVA showed that there was a statistically significant difference ($p = 0.0002$) between them, when participants selected the left screen.

Furthermore, an ANOVA was conducted between the overall success rates per image pair for each pair of labs (*Intel* vs. *UGhent*, *Intel* vs. *UWS*, *UWS* vs. *UGhent*), showing that there was a statistically significant difference ($p = 0.0046$) between the results from *UGhent* and *UWS*, there was marginally ($p = 0.0595$) no statistically significant difference between the results from *Intel* and *UGhent*, and there was no statistically significant difference ($p = 0.2618$) between the results from *Intel* and *UWS*, as also shown in Table 4.

Evaluation of the impact of scene type

As explained in "Images" section, the images used for the experiments in this study consisted of three different scene types (i.e. indoor, landscape, and night shots). In order to evaluate whether the scene type had an impact on the participants' ability to correctly detect the higher quality image, the average percentage of correct selections was computed

Table 5 Percentage of participants that selected the higher quality image per scene

Scene type	Scene	Mean	SD
Indoor	Beach	0.7810	0.1001
	Bridge	0.7951	0.0408
	Evac	0.7614	0.0218
	Flowers	0.8022	0.0777
	Fruit	0.8360	0.0628
Landscape	Autumn	0.7747	0.0813
	Ghent	0.7543	0.0313
	Green	0.8234	0.0599
	Levi	0.7755	0.0414
	Tree	0.8642	0.0303
Night	Building	0.8077	0.0250
	Corner	0.8265	0.0108
	Mirror	0.7755	0.0247
	Parking	0.7465	0.0377
	Pipes	0.7261	0.0241
All scenes		0.7900	0.0447

SD computed between the three labs

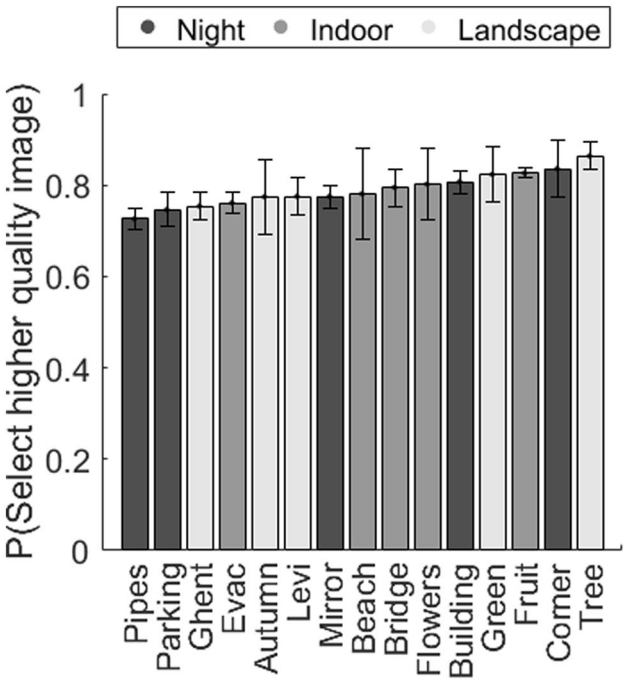

Fig. 10 Percentage of users that selected the higher quality image per scene and scene type

for each scene type and scene, as well as for all the images used. The mean correct selection percentage for each scene, as well as the standard deviations, are shown in Table 5 and Fig. 10, from where it is evident that scene type has little effect on the participants ratings. A one-way ANOVA

was also conducted between the results for the three scene types in order to establish whether there was a statistically significant difference on the participants ratings. Results showed that there was no statistically significant difference ($p = 0.8273$) between the ratings for each scene type, further supporting the argument that scene type had no significant effect on the ability of participants to correctly detect the higher quality image.

Estimation of minimum δ_{MOS} required for detecting the higher quality image

The main goal of this study is to determine the minimum MOS delta (δ_{MOS}) between two images in order for a given percentage (p_h) of people to be able to select the image with the higher quality. Each sample of the captured data consisted of the MOS delta between the two images shown to the participant along with the image and screen he/she selected. A binary flag indicating whether the subject selected the higher quality image is then computed using these data, by assigning the value 1 when the higher quality image was selected or the value 0 otherwise.

Logistic regression with binomial distribution using all the captured data was used in order to estimate the function $p_h = f(\delta_{MOS})$. The function obtained for the available MOS delta range, along with the confidence interval and the aggregated samples is depicted in Fig. 11. An estimation of the percentage of subjects that will choose photo X over photo Y for a given MOS delta between X and Y, i.e. $\delta_{MOS}(X, Y)$, can then be determined using the logistic function:

$$p_h = f\left(\delta_{MOS}\right) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \delta_{MOS})}} \tag{1}$$

where β_0 and β_1 are the β parameters computed through logistic regression and are shown in Table 6 for all the available data, as well as for each individual lab. Through the same procedure we can establish the minimum MOS delta for achieving a required probability p_h of choosing an image with a higher MOS between two images:

$$\delta_{MOS}(p_h) = \frac{logit(p_h) - \beta_0}{\beta_1} \tag{2}$$

where

$$logit(x) = -\ln\left(\frac{1}{x} - 1\right). \tag{3}$$

Furthermore, in order to evaluate the effect of the statistically significant difference in the results between the labs participating in this study, the function $p_h = f(\delta_{MOS})$ was calculated three more times by using the results from each lab separately, whereas in order to examine the effect of the statistically significant difference between the results of the three labs when users selected the left screen, the function $p_h = f(\delta_{MOS})$ was calculated two more times by using the samples from all labs for which users selected the left and the right screen respectively. The obtained functions for the individual labs and for the overall model

Table 6 β parameters, regression performance, and 95% confidence interval for the logistic regression using the data recorded by each lab

Lab	β_0	β_1	RMSE	MAE	95% CI
Intel	0.1928	1.8907	0.1436	0.1125 ± 0.1438	± 0.0232
UGhent	− 0.1578	2.8701	0.1810	0.1280 ± 0.1815	± 0.0256
UWS	0.4590	1.2593	0.1543	0.1209 ± 0.1541	± 0.0262
All	0.1678	1.9970	0.1340	0.0995 ± 0.1343	± 0.0143
All-left	0.1857	1.8198	0.1363	0.1034 ± 0.1364	± 0.0201
All-right	0.1498	2.1962	0.1424	0.1056 ± 0.1428	± 0.0214

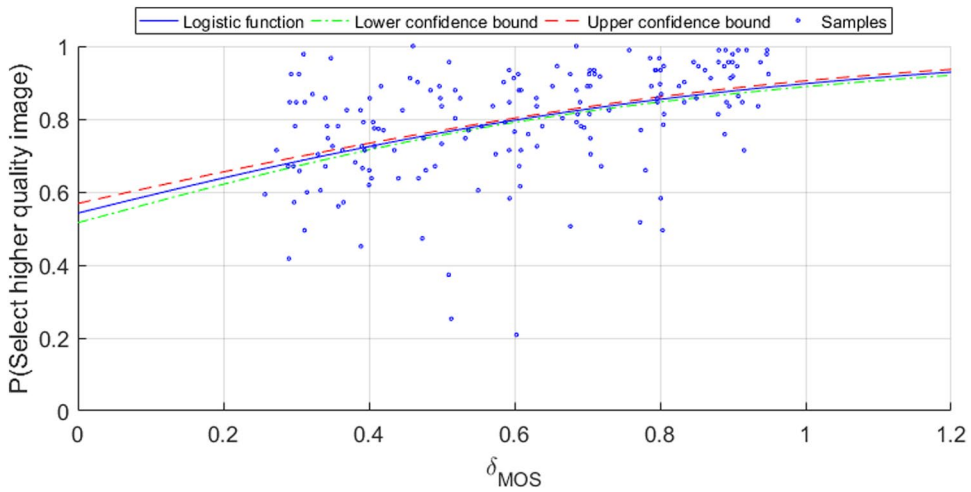

Fig. 11 The function of probability p_h of choosing an image with a higher MOS between two images, depending on the MOS delta (δ_{MOS})

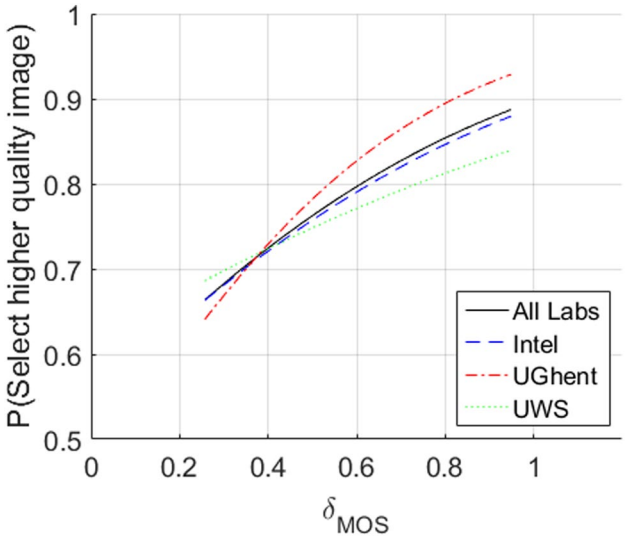

Fig. 12 The function of probability p_h of choosing an image with a higher MOS between two images, computed using the data captured from each lab

for the available MOS delta range are shown in Fig. 12, while the regression parameters for all the created models are shown in Table 6.

The minimum MOS delta required to achieve various success rates for all the examined approaches is shown in Table 7. From Table 7 it is evident that while there is a small difference in the minimum MOS delta required depending on the data used for calculating the function $p_h = f(\delta_{MOS})$, this difference can be considered as non significant since it cannot be perceived by users. For example, using the data from all the labs, the required MOS delta in order for 75% of subjects to detect the higher quality image is 0.4661 ± 0.0143. The data from each of the three labs suggest that a noticeable threshold where 75% of users can see a difference is between 0.4378 ± 0.0256 and 0.5079 ± 0.0262. For the 75% threshold, the difference of the MOS delta using all data to the MOS delta computed using the data from each individual lab is -0.0130, 0.0283, and -0.0418 for *Intel*, *UGhent*, and *UWS* respectively. Taking into consideration the confidence bounds for each prediction, the maximum difference of the model computed from all labs compared to the models computed by each individual lab becomes $\approx \pm 0.07$. Our experience shows that an absolute difference of less than 0.05 (up to 0.07 if confidence bounds are taken into consideration) in MOS value can be considered insignificant since it cannot be perceived by users. Consequently, it is evident that the statistically significant difference detected between the results from each lab (refer to "Evaluation of captured data in terms of consistency across labs" section) does not affect their overall quality and descriptive power.

Regarding the models created using the samples for which only the left or only the right screen was selected, results from the left screen suggest that a noticeable MOS delta threshold where 75% of users can see a difference is 0.5017 ± 0.0201, whereas results from the right screen set this threshold to 0.4320 ± 0.0214. It is evident that although a small difference exists between the two results (0.0697), it is not significant since it cannot be perceived by users. Furthermore, the absolute difference from the minimum MOS delta computed by the overall model is even smaller (0.0356 and 0.0341 for the left and right screen respectively), showing that the statistically significant difference detected between the results of the three labs when users selected the left screen does not significantly affect the results of the proposed model.

Discussion

From the scattered plot in Fig. 11, it is evident that significant scattering exists across the samples of this study. The MOS deltas of the image pairs were taken into consideration for creating the logistic regression model described above. As a result, different image pairs with the same MOS delta are considered as referring to the same case, thus the scatter plot in Fig. 11 depicts the aggregated image pairs. Since different image pairs are included in the same MOS delta groups, the variability stemming from these differences is not taken into consideration. In order to examine the effect of scene in the probability of selecting the higher quality image, a non-aggregated scatter plot with samples grouped by image pair was plotted in Fig. 13. It is evident that the most extreme outliers belong to different scenes. Nevertheless, 2 out of the 4 most extreme outliers ($p_h < 0.40$), which are the most extreme and the second most extreme outliers, belong to the scene "Pipes".

Table 7 Minimum MOS delta required in order for a given percentage of subjects (p_h) to detect the higher quality photo in relation to the data used

p_h	All	All-left	All-right	Intel	UGhent	UWS
0.80	0.6102 ± 0.0143	0.6598 ± 0.0201	0.5630 ± 0.0214	0.6313 ± 0.0232	0.5380 ± 0.0256	0.7363 ± 0.0262
0.75	0.4661 ± 0.0143	0.5017 ± 0.0201	0.4320 ± 0.0214	0.4791 ± 0.0232	0.4378 ± 0.0256	0.5079 ± 0.0262
0.70	0.3403 ± 0.0143	0.3636 ± 0.0201	0.3176 ± 0.0214	0.3462 ± 0.0232	0.3502 ± 0.0256	0.3083 ± 0.0262

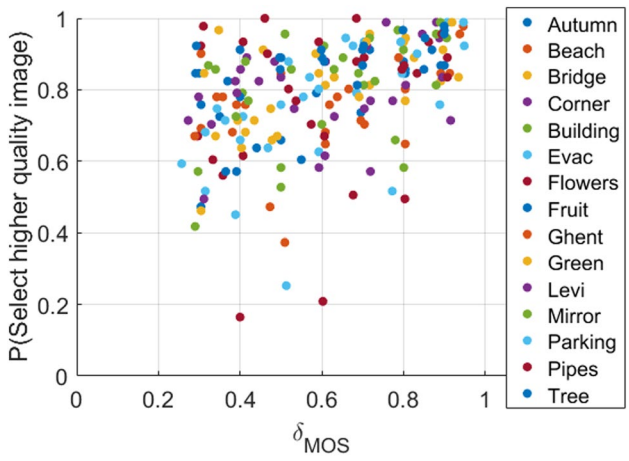

Fig. 13 Percentage of subjects that selected the higher quality image in relation to δ_{MOS} and scene

In order to better understand the problem, the pairs showing the largest difference are shown in Figs. 14 and 15. In these cases, it seems that the error comes from the crowd sourced experiment which allocated a higher score to lower quality images. For image Pipes.D (Fig. 14a) and Pipes.AA (Fig. 15a) we cannot see as much detail in the background, whereas there are obvious color errors since white colors are depicted as pink. Nevertheless, in the crowd sourced experiment, both images were given higher MOS than images Pipes.W (Fig. 14b) and Pipes.DD (Fig. 15b) respectively. The conclusion that the error may originate from the crowd sourced ratings was further supported by the MOS assigned to images Pipes.AA by *UGhent* during Phase 2 of the study, which was 3.297 compared to 3.992 assigned through crowd sourcing. Assigning a MOS of 3.297 to Pipes.AA would change the success rate of the image pair in Fig. 15 to 79.12% instead of 20.88%. Unfortunately, ratings from the other labs are not available for image Pipes.AA since it was unique to

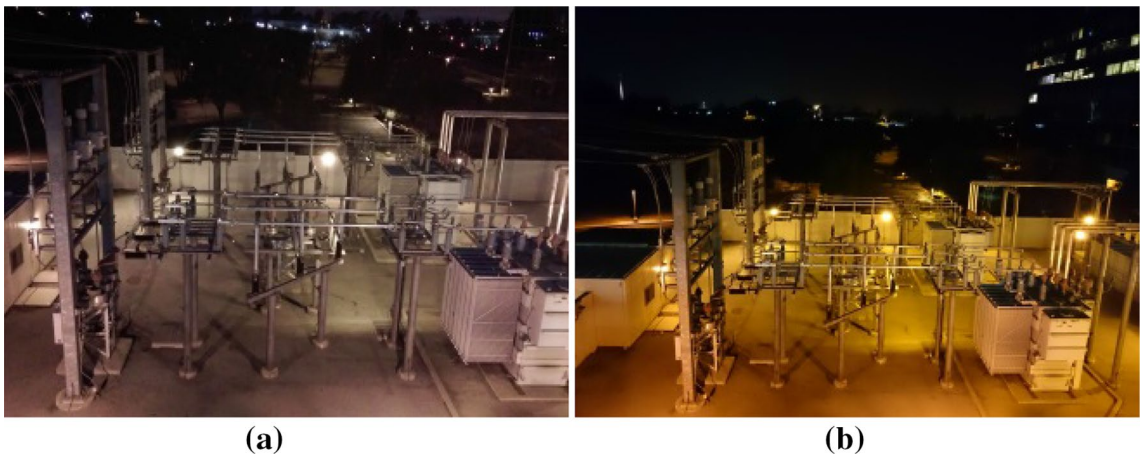

Fig. 14 The most extreme case with δ_{MOS} = 0.400 and only 16.48% of subjects choosing Pipes.D. **a** Image Pipes.D, MOS: 4.196, **b** Image Pipes.W, MOS: 3.796

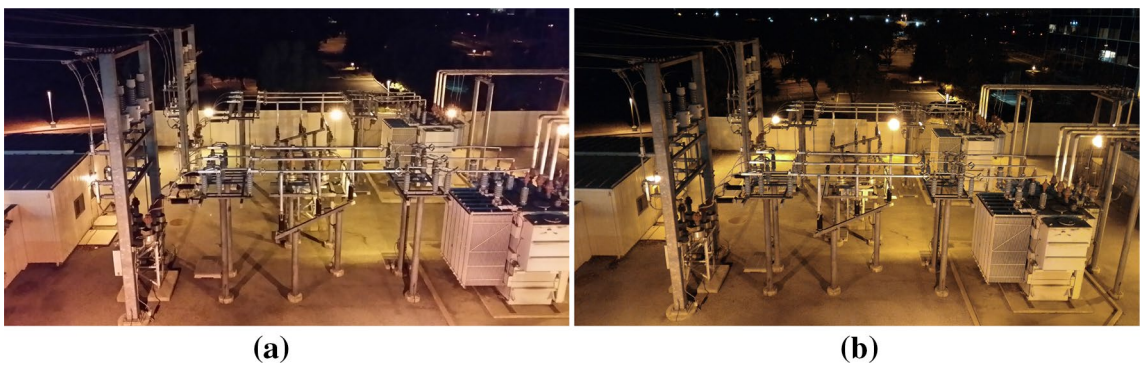

Fig. 15 The second most extreme case with δ_{MOS} = 0.603 and only 20.88% of subjects choosing Pipes.AA. **a** Image Pipes.AA, MOS: 3.992, **b** image Pipes.DD, MOS: 3.389

Fig. 16 Extremely correct case where even with small MOS difference $\delta_{MOS} = 0.311$, 97.80% of the subjects chose Pipes.BB. **a** Image Pipes.BB, MOS: 2.807, **b** Image Pipes.I, MOS: 2.496

Fig. 17 Extremely correct case where even with small MOS difference $\delta_{MOS} = 0.293$, 92.31% of the subjects chose Tree.O. **a** Image Tree.O, MOS: 4.517, **b** Image Tree.Q, MOS: 4.224

UGhent for Phase 2 of this study. Similarly, ratings for the images Pipes.D, W, and DD are not available from any lab.

Another example of large difference between the model probability and the obtained probability is presented in Fig. 16 and in Fig. 17. In these cases, even with relatively small MOS difference, only 2 out of 91 subjects made an error for the first pair ($\delta_{MOS} = 0.311$), while 7 out of 91 subjects made an error for the second pair ($\delta_{MOS} = 0.293$). For Fig. 16, image Pipes.I has evident color and focus problems, thus making it easy for subjects to select image Pipes.BB as having higher quality. In the case of Fig. 17, it seems that the slightly richer colors of image Tree.O (Fig. 17a) led most of the subjects to correctly select it. Furthermore, the MOS acquired during Phase 2 from *UGhent* for image Tree.O is similar to the one acquired through crowd sourcing (4.500 vs. 4.517). MOS ratings for the other images in Figs. 16 and 17 are not available from any lab.

The scattering of the obtained results can be caused by many different factors. In general, comparing two different images and scoring the quality are two different cognitive processes. Therefore, to obtain a more precise but less practical function given by Eq. 1, one has to use exactly the same images with different distortions. Nevertheless, the variability in image sources and capturing conditions enables the proposed model to provide a generalized "rule-of-thumb" for establishing a minimum difference between the MOS of two images in order for a certain percentage of people to be able to select the higher quality image.

It can be argued that when comparing similar non-simulated images, multiple factors apart from quality affect the users decision on which image has the higher quality. Factors like aesthetics, colors, content, etc. play an important role in users' image quality perception. Nevertheless, the use of simplified models that only take into consideration metrics that can be acquired through automated procedures is

a practical necessity. No-reference image quality tools, e.g. VIQET [18], can produce MOS ratings for images without requiring user ratings. As a result, the time needed to acquire image quality ratings is significantly reduced. Interpreting the MOS acquired by such tools poses a difficult challenge, since there is no definite analytical way to determine when the actual users will be able to tell which image has the highest quality out of group of similar images with varying MOS. In this work, we attempted to provide a solution to this problem and proposed a model that provides a MOS threshold for indicating when a certain percentage of users can detect the higher quality image. In addition, the use of a very diverse image dataset that resembles "real-world" photography use scenarios, including images acquired by multiple consumer products (phones, tablets, compact cameras, and DSLRs), increases the reliability of the acquired results and their practical usefulness.

Conclusion

In this work, the authors examined the problem of determining the minimum MOS difference required between two images in order for a given percentage of people to be able to identify the higher quality, in terms of MOS, image. A noticeable-differences subjective study was conducted by three labs, using non-simulated image stimuli created for the VIQET study, with all the images annotated with a MOS score acquired through crowd sourcing. The acquired ratings were evaluated in terms of agreement with the crowd sourced study, as well as in terms of agreement between the different labs conducting the experiments. Results showed that the acquired MOS correlated well with the crowd sourced data, while a small disagreement detected between the results for the left and the right screen of two of the labs had minimal effect on the conclusions of this study. Furthermore, while there was a small variation across different scenes in the percentage of participants that successfully detected the higher quality image from each pair, there was no statistically significant difference between the results for different scene types (indoor, landscape, night). Logistic regression was employed in order to compute the percentage of people that would successfully detect the higher quality image, as a function of the MOS difference between two images. Using this function, the minimum MOS difference required in order for a for a given percentage of people to be able to identify the higher quality image can be computed. Following this procedure, it was concluded that a MOS difference of 0.4661 ± 0.0143 is required in order for 75% of the people to be able to detect the higher quality image.

While the focus of this study was to create a generalized "rule-of-thumb" for determining a minimum MOS difference between two non-simulated images in order for the majority of users to be able to detect the higher quality image, it can be argued that other variables affect the users' perception of quality. Factors like aesthetics, artistic value, depicted content, angle of capture, color saturation, lightning, etc. also affect the decision of users regarding image quality. To this end, future work will examine the use of more complex models that will take into consideration more image characteristics than just the MOS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Ahumada AJ Jr, Peterson HA (1992) Luminance-model-based DCT quantization for color image compression. In: Proceedings of the SPIE, vol 1666, pp 365–374. https://doi.org/10.1117/12.135982
2. Deng Y, Zhang Y, Yang D, Chen Z (2017) Towards foveated just noticeable difference modeling for virtual reality. Electron Imaging 12:198–201. https://doi.org/10.2352/ISSN.2470-1173.2017.12.IQSP-243
3. Ferzli R, Karam LJ (2009) A no-reference objective image sharpness metric based on the notion of just noticeable blur (JNB). IEEE Trans Image Process 18(4):717–728. https://doi.org/10.1109/TIP.2008.2011760
4. Hontsch I, Karam LJ (2002) Adaptive image coding with perceptual distortion control. IEEE Trans Image Process 11(3):213–222. https://doi.org/10.1109/83.988955
5. ITU-R (1998) Subjective assessment methods for image quality in high-definition television. ITU-R Rec. BT.710-4
6. ITU-R (2012) Methodology for the subjective assessment of the quality of television pictures. ITU-R Rec. BT.500-13
7. ITU-R (2015) Parameter values for the HDTV standards for production and international programme exchange. ITU-R Rec. BT.709-6
8. Jazi SD, Heath M (2017) The spatial relations between stimulus and response determine an absolute visuo-haptic calibration in pantomime-grasping. Brain Cognit 114:29–39. https://doi.org/10.1016/j.bandc.2017.03.002
9. Jia Y, Lin W, Kassim AA (2006) Estimating just-noticeable distortion for video. IEEE Trans Circuits Syst Video Technol 16(7):820–829. https://doi.org/10.1109/TCSVT.2006.877397
10. Kamble V, Bhurchandi K (2015) No-reference image quality assessment algorithms: a survey. Opt Int J Light Electron Opt 126(1112):1090–1097. https://doi.org/10.1016/j.ijleo.2015.02.093
11. Kim M, Song KS, Kang MG (2017) No-reference image contrast assessment based on just-noticeable-difference. Electron Imaging

12:26–29. https://doi.org/10.2352/ISSN.2470-1173.2017.12.IQSP-221
12. Lin JY, Jin L, Hu S, Katsavounidis I, Li Z, Aaron A, Kuo CCJ (2015) Experimental design and analysis of JND test on coded image/video, vol 9599, pp 95,990Z–95,990Z-11. https://doi.org/10.1117/12.2188389
13. Liu A, Lin W, Paul M, Deng C, Zhang F (2010) Just noticeable difference for images with decomposition model for separating edge and textured regions. IEEE Trans Circuits Syst Video Technol 20(11):1648–1652. https://doi.org/10.1109/TCSVT.2010.2087432
14. Peterson HA, Ahumada AJ Jr, Watson AB (1993) Improved detection model for DCT coefficient quantization, vol 1913, pp 191–201. https://doi.org/10.1117/12.152693
15. Saad MA, Pinson MH, Nicholas DG, Kets NV, Wallendael GV, Silva RD, Jaladi RV, Corriveau PJ (2015) Impact of camera pixel count and monitor resolution perceptual image quality. In: 2015 colour and visual computing symposium (CVCS), pp 1–6. https://doi.org/10.1109/CVCS.2015.7274887
16. Saad MA, McKnight P, Quartuccio J, Nicholas D, Corriveau RJP (2016) Online subjective testing for consumer-photo quality evaluation. J Electron Imaging 25(4):043009. https://doi.org/10.1117/1.JEI.25.4.043009
17. Tran TD, Safranek R (1996) A locally adaptive perceptual masking threshold model for image coding. In: 1996 IEEE international conference on acoustics, speech, and signal processing conference proceedings, vol 4, pp 1882–1885. https://doi.org/10.1109/ICASSP.1996.544817
18. VQEG (2015) Video quality experts group: progress report 2015 v.1
19. Wang S, Ma L, Fang Y, Lin W, Ma S, Gao W (2016) Just noticeable difference estimation for screen content images. IEEE Trans Image Process 25(8):3838–3851. https://doi.org/10.1109/TIP.2016.2573597
20. Watson A (1993) DCTune: a technique for visual optimization of DCT quantization matrices for individual images. In: Society for information display (SID) digest, vol 24, pp 946–949
21. Wei Z, Ngan KN (2009) Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain. IEEE Trans Circuits Syst Video Technol 19(3):337–346. https://doi.org/10.1109/TCSVT.2009.2013518
22. Wu J, Shi G, Lin W, Liu A, Qi F (2013) Just noticeable difference estimation for images with free-energy principle. IEEE Trans Multimed 15(7):1705–1710. https://doi.org/10.1109/TMM.2013.2268053
23. Wu J, Li L, Dong W, Shi G, Lin W, Kuo CCJ (2017) Enhanced just noticeable difference model for images with pattern complexity. IEEE Trans Image Process. https://doi.org/10.1109/TIP.2017.2685682
24. Yang K, Wan S, Wu HR, Lin W, Tan D, Gong Y, Xie L (2016) Detection and estimation of supra-threshold distortion levels of pictures based on just-noticeable difference. In: 2016 visual communications and image processing (VCIP), pp 1–4. https://doi.org/10.1109/VCIP.2016.7805539
25. Yang X, Lin W, Lu Z, Ong E, Yao S (2005) Motion-compensated residue preprocessing in video coding based on just-noticeable-distortion profile. IEEE Trans Circuits Syst Video Technol 15(6):742–752. https://doi.org/10.1109/TCSVT.2005.848313
26. Yang X, Ling W, Lu Z, Ong E, Yao S (2005) Just noticeable distortion model and its applications in video coding. Signal Process Image Commun 20(7):662–680. https://doi.org/10.1016/j.image.2005.04.001
27. Young GW, Murphy D, Weeter J (2017) Haptics in music: the effects of vibrotactile stimulus in low frequency auditory difference detection tasks. IEEE Trans Haptics 10(1):135–139. https://doi.org/10.1109/TOH.2016.2646370
28. Zhang X, Lin W, Xue P (2005) Improved estimation for just-noticeable visual distortion. Signal Process 85(4):795–808. https://doi.org/10.1016/j.sigpro.2004.12.002
29. Zhang X, Lin W, Xue P (2008) Just-noticeable difference estimation with pixels in images. J Vis Commun Image Represent 19(1):30–41. https://doi.org/10.1016/j.jvcir.2007.06.001
30. Zhang X, Wang S, Gu K, Lin W, Ma S, Gao W (2017) Just-noticeable difference-based perceptual optimization for JPEG compression. IEEE Signal Process Lett 24(1):96–100. https://doi.org/10.1109/LSP.2016.2641456